# 融合项目偏差与用户偏好的推荐算法 *

程　磊，高茂庭

(上海海事大学 信息工程学院，上海 201306)

**摘　要：**针对协同过滤推荐中由于项目和用户间关联因素的相互影响而存在项目偏差和用户偏好的问题，提出一种融合项目偏差与用户偏好的推荐算法。先进行聚类处理，包括 LDA 主题建模生成项目簇和 K-means 聚类生成用户簇；再依次根据项目簇和用户簇的约束生成项目偏差分，同时以用户项目评分及项目类型为基础，经过概率转移得到用户偏好分；最后以项目簇内已有评分的均值为基础，对项目偏差分和用户偏好分进行线性加权生成预测评分。对比实验表明，新算法能够根据不同的近邻得到合理的推荐，提高推荐的准确度。

**关键词：**协同过滤；主题建模；聚类；项目偏差；用户偏好

**中图分类号：**TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.05.0298

## Recommendation algorithm combining item deviation and user preference

Cheng Lei, Gao Maoting

(*College of Information Engineering, Shanghai Maritime University, Shanghai* 201306, *China*)

**Abstract:** Aiming at the problem that there exists item deviation and user preferences in collaborative filtering recommendation for the interaction between factors related in items and users, this paper proposed a recommendation algorithm integrated item deviation and user preference. Firstly it clustered to generate item clusters on LDA topics modeling and to get user clusters by using K-means; then it generated item deviation score on the constraints of item cluster and user cluster, and obtained user preference score with probability transfer on user-item score and item type. Finally it weighted the item deviation score and user preference score linearly to form the prediction score based on the existing scoring average in the item cluster. Comparison experiments show that the new algorithm could obtain reasonable recommendation based on different neighbors and improve recommendation accuracy.

**Key words:** collaborative filtering; topic modeling; clustering; item deviation; user preference



## 0　引言

推荐系统作为一种是帮助用户快速选择有效信息的重要工具，正在被越来越多的电子商务和社交网站用来改善用户体验。推荐算法主要包括基于协同过滤的算法、基于内容的算法和基于标签的算法[1]。作为常用的推荐算法，协同过滤主要依据用户—项目评分信息进行评分预测，存在一些不足使算法准确度难以提高，一方面，由于受项目间关联因素的相互影响，实际存在项目偏差问题；另一方面，却缺少考虑不同用户对各项目类型的偏好因素。

为此，Shi 等人[2]提出一种基于新型概率主题模型，在相似度计算时引入两个项目之间相异性的惩罚项，减少不相关的项目对于准确度的影响。Qiao 等人[3]提出一种结合用户属性和项目内容的推荐算法，利用 LDA(latent Dirichlet allocation)模型分别对用户属性和项目内容进行主题分析，挖掘出用户对项目的偏好。Zhao 等人[4]提出一种基于特征转移和概率矩阵分解的推荐算法，将信任矩阵集成到评分矩阵中，用户的评分只受到自身属性以及信任的人影响，从而过滤掉无关用户。Zhou 等人[5]提出一种评估协同过滤的 LDA 模型，通过给主题模型添加用户评分信息来进行协同过滤，并利用用户对于项目的偏好给出合理的推荐。原福永[6]等人提出一种基于项目的协同过滤算法，将项目分类和 K 近邻引入到 Slope One 算法，从而过滤掉不相关的项目。刘慧婷等人[7]提出一种基于用户偏好的矩阵分解算法，通过用户项目评分矩阵和矩阵分解得到的项目属性矩阵计算用户的偏好，提高了预测的准确度。

这些研究一定程度上考虑了项目偏差或用户偏好因素对预测的影响，取到了较好的成效，基于此，提出一种融合项目偏差和用户偏好的推荐算法（item deviation and user preference combination filtering，IUCF），通过对项目和用户分别进行聚类处理，利用最近邻协同过滤和概率转移挖掘项目偏差和用户偏

**收稿日期：**2018-05-10；**修回日期：**2018-06-15　**基金项目：**国家自然科学基金资助项目(61202022)；上海海事大学研究生创新基金资助项目(2017ycx061)

**作者简介：**程磊（1994-），男，山东潍坊人，硕士研究生，主要研究方向为数据挖掘、推荐算法（email_chenglei@163.com）；高茂庭（1963-），男，江西九江人，教授，博士，主要研究方向为智能信息处理、数据库与信息系统.

好，并将两者融入用户对项目的预测评分中完成推荐。

## 1 相关研究

### 1.1 基于用户的协同过滤算法

基于用户的协同过滤算法分为以下几步：a)根据用户-项目评分矩阵计算用户之间的相似度；b)根据相似度选取近邻用户；c)根据近邻用户给出预测分数。相似度的计算采用皮尔逊相关系数[8]：

$$Sim(a,b)=\frac{\sum_{n\in I_{ab}}(r_{an}-\overline{r_a})(r_{bn}-\overline{r_b})}{\sqrt{\sum_{n\in I_{ab}}(r_{an}-\overline{r_a})^2}\sqrt{\sum_{n\in I_{ab}}(r_{bn}-\overline{r_b})^2}} \tag{1}$$

其中: $I_{ab}$ 表示用户 $a$ 与用户 $b$ 共同评分过的项目集合，$r_{an}$ 和 $r_{bn}$ 分别表示用户 $a$ 和用户 $b$ 对项目 $n$ 的实际评分，$\overline{r_a}$ 和 $\overline{r_b}$ 分别表示用户 $a$ 和用户 $b$ 所有评分的平均值。

### 1.2 LDA 主题模型

LDA 主题模型[9]首先使用 Dirichlet 概率分布来设置文档的潜在概率，然后使用抽样算法来估计文档-主题概率分布和主题-词汇概率分布。抽样算法采取 Gibbs 采样[10]：

$$p(z_i=k\mid \vec{z}_{\neg i},\vec{w})\propto \frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(k)}+\alpha_k)}\bullet\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)}+\beta_t)} \tag{2}$$

其中: $p(z_i=k\mid \vec{z}_{\neg i},\vec{w})$ 表示排除第 $i$ 个词汇，根据文档集 $\vec{w}$ 中其它词汇序列的主题分布来计算第 $i$ 个词汇属于第 $k$ 个主题的概率，其中 $z_i$ 表示语料库 $\vec{z}$ 中的第 $i$ 个词汇对应的主题；$n_{m,\neg i}^{(k)}$ 表示排除第 $i$ 个词汇，第 $m$ 篇文档中主题 $k$ 的词汇次数；$n_{k,\neg i}^{(t)}$ 表示排除第 $i$ 个词汇，第 $k$ 个主题中词汇 $t$ 的次数；$\alpha_k$ 和 $\beta_t$ 分别表示文档-主题分布和主题-词汇分布的 Dirichlet 先验参数。

### 1.3 问题描述与分析

在进行评分预测过程中，由于受项目间关联因素的相互影响，实际存在项目偏差问题，例如动作题材电影的评分计算时，除同一题材电影的评分外，许多无关题材的评分也参与了计算，导致预测评分偏离了实际；同样，在近邻选取时，往往是在整个用户集里找寻，并未充分考虑用户的内在属性，无法准确体现不同用户对不同类型项目所存在的一些偏好，例如：不同年龄段人群间爱好存在一定的差异，表面上打分相近用户，实际上可能有较大差距。同时，用户的偏好也影响用户对项目的评分，例如：当电影同时存在用户喜欢的动作题材和不喜欢的恐怖题材，用户对电影的打分就会受到其个人偏好的影响。需要准确挖掘项目偏差和用户偏好，以对最终预测评分进行纠偏。

针对项目偏差挖掘，文献[6]分别从用户和项目的角度对项目偏差进行过滤，最后再进行混合加权，但是这种方式计算量比较大。文献[11]使用用户间信任度与相似度的线性加权作为最近邻的选取依据，但其在计算时依然未考虑项目的内在属性。针对用户偏好挖掘，文献[7]通过矩阵分解算法挖掘出用户的偏好，但矩阵分解算法可解释性差。文献[12]简单使用包括某类型的项目评分的均值表示类型的评分，导致用户偏好区分度不高。

为此，在 IUCF 算法中，充分考虑项目和用户的内在属性，通过对项目类型和用户属性进行聚类，使得项目偏差更加准确可靠。通过统计项目类型被标记次数的比例，再结合用户评分矩阵，使得出现频率高的类型获得较高的打分，从而使得到的用户偏好可靠。

## 2 融合项目偏差与用户偏好的推荐算法

IUCF 算法中融入两部分内容：基于项目簇和用户簇生成项目偏差分和基于项目类型生成用户偏好分，并以目标用户所在项目簇的已有评分均值为基础，对两部分线性加权生成最终预测评分，算法模型如图 1 所示。

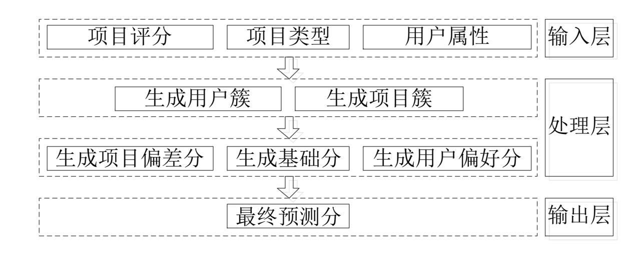



图 1 IUCF 算法模型

### 2.1 聚类处理

为准确地体现项目偏差，需要在项目偏差中同时考虑项目类型和用户属性；同时为准确地挖掘用户偏好，在计算用户偏好时考虑用户对不同项目类型的喜好程度。为此，通过对项目类型和用户属性的聚类，生成项目簇和用户簇，计算项目偏差和用户偏好，从而提高推荐的准确度。

#### 2.1.1 LDA 主题建模生成项目簇

LDA 主题建模是一种文档主题生成模型，能够识别语料库中潜在的主题信息。在模型中，每一篇文档看作是由许多主题所构成，而每一个主题又由许多词汇所构成。将所有的项目类型作为文档，将类型作为词汇，找寻每个词汇所在的主题。设 $I$ 为项目集合，$I=\{I_1,I_2,\cdots,I_n\}$，其中，$I_n$ 表示第 $n$ 个项目。定义项目类型 $I_n=(i_1,i_2,\cdots,i_m)$，其中，$i_m$ 表示项目 $I_n$ 的第 $m$ 个类型，例如，当 $I_n$ =（Action, Drama, War）时，表示项目 $I_n$ 的类型为 Action、Drama 和 War。

将项目集合进行 LDA 主题建模，使用 Gibss 采样法，得到项目-类型主题分布 $\hat{z}$ 和主题-类型概率矩阵 $D_{t\times m}$，其中主题数目需要根据 LDA 聚类后算法的准确度确定。

在 $\hat{z}$ 中，每个项目 $I_n$ 的类型 $i_m$ 都有一个主题标号，形式如 $I_n=(i_1:z_1,i_2:z_2,\cdots,i_m:z_t)$，其中，$i_m:z_t$ 表示 $i_m$ 对应的主题标号为 $z_t$，例如，Action:3 表示类型 Action 属于主题 3。

根据 $\hat{z}$ 建立项目-主题隶属矩阵 $\hat{I}_{n\times t}$，当项目的类型 $i_m$ 属于主题 $t$ 时，$\hat{i}_{nt}$ =1，否则，$\hat{i}_{nt}$ =0，如式（3）所示。

$$\hat{I}_{n\times t}=\begin{bmatrix}\hat{i}_{11} & \cdots & \hat{i}_{1t}\\ \vdots & & \vdots\\ \hat{i}_{n1} & \cdots & \hat{i}_{nt}\end{bmatrix} \tag{3}$$

由项目—主题隶属矩阵的任意一行可以得到项目主题簇，记为 $\mathrm{CN}_n$；由项目—主题隶属矩阵的任意一列，可以得到主题项目簇，记为 $\mathrm{CT}_t$。如式（4）（5）所示。

$$\mathrm{CN}_n=\{i\,|\,\hat{i}_{ni}=1,i\in[1,t]\} \tag{4}$$

$$\mathrm{CT}_t=\{j\,|\,\hat{i}_{jt}=1,j\in[1,n]\} \tag{5}$$

其中：$\mathrm{CN}_n$ 表示项目 $n$ 所属的主题集合，$\mathrm{CT}_t$ 表示主题 $t$ 包含的项目集合，$i$ 表示主题号，$j$ 表示项目号。

对于目标项目 $n$，先找到其对应的 $\mathrm{CN}_n$，然后找到每个主题所对应的 $CT_t$，最后将所有的 $\mathrm{CT}_t$ 进行并集运算得到目标项目 $n$ 所在的项目簇 $C_n$，如式（6）所示。

$$C_n=\bigcup_{t\in\mathrm{CN}_n}\mathrm{CT}_t \tag{6}$$

2.1.2 K-means 聚类生成用户簇

在当前的网络环境下，可以准确获取用户属性，用户属性包括用户的性别，年龄和职业。定义用户集合为 $Q=\{Q_1,Q_2,\cdots,Q_i\}$，用户属性为 $Q_i=(q_1,q_{2,\ldots},q_k)$，其中，$Q_i$ 表示第 $i$ 个用户，$q_k$ 表示用户 $i$ 的第 $k$ 个基本属性，例如，当 $Q_i$ =（男，23，teacher）时，表示用户 $Q_i$ 的性别为男，年龄为 23，职业为 teacher。

为了进行 K-means 聚类，需要对用户的基本信息进行预处理，采用数字编码[1-9]的方式对用户的基本信息进行预处理。针对性别，将男女分别编码为 1 和 2；针对年龄，根据文献[13]所提出的 5 组划分方式，将用户年龄划分为少儿组（0~19 岁），青年组（20~39 岁），壮年组（40~59 岁），实年组（60~79 岁）和老年组（≥80 岁），并依次编码为 1,2,3,4,5。针对职业，文献[14]将二八定理用于信息评估，定理指出：任何一组东西，最重要的东西只占大约 20%，其余的只占 80%。同理，统计所有用户的职业种类及每种职业的用户数量，根据每种职业的用户数量对职业进行降序排名，对排在前 20%的职业给予单独编码，剩余的职业归为一类。例如，统计 MovieLens 中用户的职业情况，根据统计，一共 21 个职业，取用户数量前 4 的职业进行单独编码，剩余的职业归为一类，所以用户的职业编码依次为 1,2,3,4,5。经过数字编码后，用户属性表示为数字编码，例如，$Q_i$ =（男，23，teacher）表示为 $Q_i$ =（1,1,5）。

由于 K-means 聚类是一种无监督学习方法，具体的聚类个数需要根据所有类簇的误差平方和(sum of squares errors, SSE)来确定，采用肘方法[15],该方法选择簇内误差平方和关于簇数曲线的拐点作为聚类数。经过 K-means 聚类后得到用户 $a$ 所在的用户簇 $U_a$，如式（7）所示。

$$U_a=\{u_j\,|\,u_j\in Q,j\in[1,i]\} \tag{7}$$

其中：$u_j$ 表示用户簇中的第 $j$ 个用户，$Q$ 表示用户集合。

## 2.2 计算项目偏差分

采用目标项目被打分时的偏离项目均值的程度来衡量项目偏差，为了准确地表达项目偏差分，主要通过项目簇和用户簇对项目和用户进行过滤，计算每个项目与其所在项目簇已有评分的均值的差值。其计算过程如下：

a）在项目簇内计算增强的评分相似度。

在项目簇 $C_n$ 内，根据用户-项目评分矩阵计算用户之间的增强评分相似度 $\hat{S}im(a,b)$，如式（8）所示。

$$\hat{S}im(a,b)=\frac{\sum_{n\in I_{ab}}(r_{an}-\bar{r}_a^n)(r_{bn}-\bar{r}_b^n)}{\sqrt{\sum_{n\in I_{ab}}(r_{an}-\bar{r}_a^n)^2}\sqrt{\sum_{n\in I_{ab}}(r_{bn}-\bar{r}_b^n)^2}} \tag{8}$$

其中：$\bar{r}_a^n$ 和 $\bar{r}_b^n$ 分别表示 $r_{an}$ 和 $r_{bn}$ 各自项目对应的 $C_n$ 中的已有评分的均值。

b）在用户簇内求出目标用户的最近邻。

生成目标用户与其他用户的增强相似度，在用户簇 $U_a$ 内，选取与目标用户 $a$ 相关系数最高的前 $k$ 个用户组成目标用户的最近邻 $\hat{N}(a)_k$，如式（9）所示。

$$\hat{N}(a)_k=\{b_j\,|\,b_j\in Desc(\hat{S}im(a,b_j),b_j\in U_a),j\in[1,k]\} \tag{9}$$

其中：$b_j$ 为与用户 $a$ 相关系数从高到低的第 $j$ 个用户，$Desc(\hat{S}im(a,b_j),b_j\in U_a)$ 为在用户簇 $U_a$ 内与用户 $a$ 的相似度由高到低的排序序列。

c）采用加权平均偏差生成项目偏差分。

在得到目标用户的最近邻后，采用加权平均偏差作为用户簇内用户 $a$ 对目标项目 $n$ 的项目偏差分，如式（10）：

$$\mathrm{ID}_{an}=\frac{\sum_{b\in\hat{N}(a)_k}\hat{S}im(a,b)(r_{bn}-\bar{r}_b^n)}{\sum_{b\in\hat{N}(a)_k}\hat{S}im(a,b)} \tag{10}$$

其中：$\mathrm{ID}_{an}$ 为用户簇内用户 $a$ 对目标项目 $n$ 的项目偏差分。

## 2.3 计算用户偏好分

项目偏差分是基于相似度生成的，但由于其受限于共同评分项，当共同评分项很少甚至不存在时，项目偏差分就失去了调节意义，因此，考虑计算用户偏好分，它反映了用户对不同项目的偏好程度。算法先从项目类型偏好分出发，挖掘用户的类型喜好，再计算主题偏好分，最后计算用户偏好分。其计算过程如下：

a)由用户类型偏好分生成用户类型喜好。

用户类型喜好是基于用户类型偏好分的，用户类型偏好分表示不同类型在用户评分总和中所占的分值，即不同类型对总分所起的贡献比例，如式（11）所示。

$$p_{ai}=\frac{s_{ai}}{\sum_{i=1}^{m}s_{ai}}\sum_{j=1}^{n}r_{aj} \tag{11}$$

其中：$p_{ai}$ 表示用户 $a$ 对类型 $i$ 偏好分，$s_{ai}$ 表示用户 $a$ 对类型 $i$ 的评论次数，$r_{aj}$ 表示用户 $a$ 对项目 $j$ 的评分，$m$ 表示类型数，$n$ 表示项目数。$s_{ai}$ 是用户-类型评次矩阵 $S_{a\times m}$ 的每一项，$S_{a\times m}$ 通过用户—项目隶属矩阵 $\hat{R}_{a\times n}$ 和项目-类型隶属矩阵 $\hat{I}_{n\times m}$ 的对应项相乘得到。在 $\hat{R}_{a\times n}$ 中，当用户 $a$ 存在对 $I_n$ 的评分时，$\hat{r}_{an}=1$，否则 $\hat{r}_{an}=0$。同理，在 $\hat{I}_{n\times m}$ 中，当 $\hat{I}_{n\times m}$ 中存在属于类型 $m$ 的类型时，$\hat{i}_{nm}=1$，否则 $\hat{i}_{nm}=0$。如式（12）：

$$S_{a\times m}=\hat{R}_{a\times n}\times\hat{I}_{n\times m} \tag{12}$$

用户的类型喜好存在喜欢和不喜欢，z-score 标准化反映了数值与均值的差异程度，其结果的正负刚好用来反映用户对不同类型的喜好，因此对用户的类型偏好分进行 z-score 归一化，如式（13）所示。

$$\hat{p}_{ai}=\frac{p_{ai}-\mu_a}{\sigma_a} \tag{13}$$

其中：$\hat{p}_{ai}$ 表示用户 $a$ 对类型 $i$ 的喜好，$\mu_a$ 表示用户 $a$ 的所有类型偏好分的均值，$\sigma_a$ 表示用户 $a$ 的所有类型偏好分的标准差。

b)依次计算主题偏好分和用户偏好分。

先计算用户对主题的偏好分，主题偏好分反映了用户对于主题的偏好程度，它是用户-类型喜好矩阵 $\hat{P}_{a\times m}$ 和类型-主题概率矩阵 $D_{m\times t}$ 对应项相乘得到，计算形式同式（12），其中 $\hat{P}_{a\times m}$ 由 $\hat{p}_{ai}$ 组成，$D_{m\times t}$ 是 $D_{t\times m}$ 的转置矩阵；再将同一项目主题簇 $\mathrm{CN}_n$ 中的用户主题偏好分累加；最后采用平均法生成用户 $a$ 对项目 $n$ 的偏好分 $\mathrm{UP}_{an}$。如式（14）所示。

$$\mathrm{UP}_{an}=\frac{\sum_{t\in\mathrm{CN}_n}(\sum_{i=1}^{m}\hat{p}_{ai}d_{it})}{N_n} \tag{14}$$

其中：$d_{it}$ 表示 $D_{m\times t}$ 中类型 $m$ 对应的主题 $t$ 的概率，$N_n$ 表示项目 $n$ 所属 $\mathrm{CN}_n$ 中的主题个数。

### 2.4　生成最终预测评分

目标项目的预测评分是以目标项目所在的项目簇的已有评分的均值为基础，添加项目偏差分和用户偏好分而生成的。通过权重系数 $\lambda$ 对两种预测分进行调节，从而得到用户 $a$ 对项目 $n$ 的最终预测评分 $T_{an}$，如式（15）所示。

$$T_{an}=\overline{r}_a^n+\lambda\mathrm{ID}_{an}+(1-\lambda)\mathrm{UP}_{an} \tag{15}$$

其中：$\overline{r}_a^n$ 表示用户 $a$ 所预测项目 $n$ 的项目簇 $C_n$ 中的已有评分的均值。最终的预测评分受到权重系数 $\lambda$ 的影响，当 $\lambda=0$ 时，预测评分只受到用户偏好分的调节，当 $\lambda=1$ 时，预测评分只受到项目偏差分的调节，具体的 $\lambda$ 值需要根据实验给出。

### 2.5　IUCF 的算法描述

算法　融合项目类型与用户属性的推荐算法（IUCF）

输入：用户-项目评分矩阵 $R_{a\times n}$，项目集合 $I$，用户集合 $Q$，权重系数 $\lambda$。

输出：目标项目的预测评分 $T_{an}$。

a)先对 $I$ 进行 LDA 主题建模得到项目-类型主题分布 $\hat{Z}$ 和主题-类型概率矩阵 $D_{t\times m}$，再根据 $\hat{Z}$ 建立项目-主题隶属矩阵 $\hat{I}_{n\times t}$，随后由 $\hat{I}_{n\times t}$ 生成项目主题簇 $\mathrm{CN}_n$ 和主题项目簇 $\mathrm{CT}_t$，最后由 $\mathrm{CN}_n$ 和 $\mathrm{CT}_t$ 生成项目簇 $C_n$。

b)先对 $Q$ 中的每个用户 $Q_i$ 的基本信息进行数字编码，再对编码后的 $Q$ 进行 k-means 聚类，最后生成用户簇 $U_a$。

c)先在 $C_n$ 内根据 $R$ 使用皮尔逊相关系数计算用户间的增强评分相似度 $\hat{S}im(a,b)$，再在 $U_a$ 内根据 $\hat{S}im(a,b)$ 由高到低生成目标用户的最近邻 $\hat{N}(a)_k$，最后采取加权平均偏差法生成项目偏差分 $\mathrm{ID}_{an}$。

d)先对用户-项目隶属矩阵 $\hat{R}_{a\times n}$ 和项目-类型隶属矩阵 $\hat{I}_{n\times m}$ 进行矩阵相乘得到用户-类型评次矩阵 $S_{a\times m}$，再由 $S_{a\times m}$ 和 $R_{a\times n}$ 的对应项生成用户类型偏好分 $p_{ai}$，最后对 $p_{ai}$ 进行 z-score 标准化得到用户类型喜好 $\hat{p}_{ai}$。

e)先由 $\hat{p}_{ai}$ 建立用户-类型喜好矩阵 $\hat{P}_{a\times m}$，将 $D_{t\times m}$ 转置生成类型-主题概率矩阵 $D_{m\times t}$，再根据 $\hat{P}_{a\times m}$ 和 $D_{m\times t}$ 对应项相乘生成用户主题偏好分，再累加同一 $\mathrm{CN}_n$ 中的用户主题偏好分，最后采用平均法生成用户偏好分 $\mathrm{UP}_{an}$。

f)以目标用户所属 $C_n$ 的已有评分的均值 $\overline{r}_a^n$ 为基础，通过权重系数 $\lambda$ 对 $\mathrm{ID}_{an}$ 和 $\mathrm{UP}_{an}$ 进行调节，最终生成 $T_{an}$

### 2.6　IUCF 的时间复杂度分析

设项目数为 $n$，项目类型数为 $m$，主题数为 $t$，Gibbs 迭代次数为 $g$，用户数为 $a$，用户的基本信息数为 $q$，k-means 聚类个数为 $s$，k-means 迭代次数 $d$，最近邻个数为 $k$。在用户数和项目数保持一定的情况下，算法分为离线部分和在线部分。

离线部分：步骤 a)：LDA 主题建模生成项目簇的时间复杂度为 $O(gnmt)$，步骤 b)：k-means 聚类生成用户簇的时间复杂度为 $O(daqs)$，步骤 c)：在项目簇内计算增强的评分相似度的时间杂度为 $O(a^2n)$，步骤 d)：由用户类型偏好分生成用户类型喜好的时间复杂度为 $O(anm)$。

在线部分：步骤 c)：在用户簇内求出目标用户的最近邻和采用加权平均偏差生成项目偏差分的时间复杂度为 $O(a\log_2 a+ka)$，步骤 d)：由主题偏好分生成用户偏好分的时间复杂度为 $O(amt)$，步骤 e)：生成最终评分的时间复杂度为 $O(1)$。

综上，离线部分包含 ac 和 bd 两个独立的分支，而 LDA 主题建模的时间复杂度远大于 k-means，所以时间复杂度为

$O(gnmt+a^2n)$，在线部分时间复杂度为$O(a\log_2 a+amt)$。

## 3　实验结果与分析

### 3.1　实验数据集和实验环境

实验数据集采用美国明尼苏达大学的 GroupLens 小组开发并维护的 MovieLens 100K 数据集，包含 943 个用户对 1682 部电影的 10 万多条评分，以及电影类型和用户属性等内容。将数据集的 80%作为训练集，剩余的 20%作为测试集，采用五折交叉验证的方式进行实验。

实验环境为 Intel Core i5 处理器和 8G 内存，Windows7 x64 操作系统，算法使用 Python3.5 语言实现。

### 3.2　实验度量标准

准确性采用平均绝对误差（Mean Absolute Error, MAE）作为度量标准，它可以直观反映推荐质量的高低，MAE 越小，证明推荐准确度越好，如式（16）所示。

$$\mathrm{MAE}=\frac{\sum_{n=1}^{N}|T_{an}-r_{an}|}{|\mathrm{N}|} \tag{16}$$

其中：$T_{an}$ 表示用户 $a$ 对项目 $n$ 的预测评分，$r_{an}$ 表示用户 $a$ 对项目 $n$ 的实际评分，N 表示预测的项目数。

### 3.3　算法相关参数确定实验

算法中需要确定的参数包括 K-means 聚类个数、LDA 聚类主题数目和权重系数 $\lambda$。

#### 3.3.1 K-means 聚类个数的确定

为了确定合适的 k-means 聚类数，采用肘部方法，依次选取聚类的个数为 2、3、4、5、6、7、8，分别计算 SSE，结果如图 2 所示。

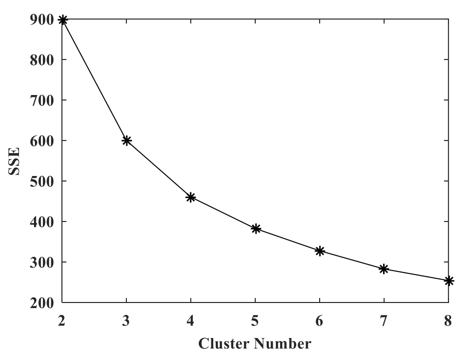



图 2　不同聚类个数下的 MAE 值变化

图 2 中，随着聚类数目的增加，SSE 的值依次下降。从图中可以看出，当聚类个数从 2 到 3 时，SSE 的值下降了大约 300，而聚类个数从 3 开始，SSE 的值下降速度明显小于 300，同时 3 所对应的点也刚好为肘方法的拐点，因此，当聚类个数为 3 时，聚类效果最好。

#### 3.3.2 LDA 聚类主题数目的确定

为了确定合适的 LDA 聚类的主题数目，设置 $\lambda$=1，即仅通过项目偏差分的算法去确定最佳主题数目，固定近邻数目为 10、30、50，采用 Gibbs 抽样方法，依次选取主题数目为 12、13、14、15、16、17、18 进行实验，同时根据文献[16]的实践进行参数设置，$\alpha=50/T$，$\beta=0.01$，如图 3 所示。

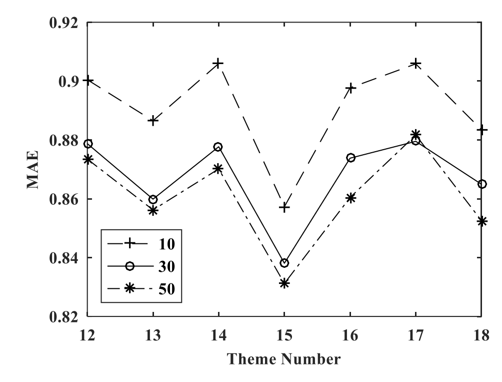



图 3　不同主题数下的 MAE 值变化

图 3 中，在近邻数目为 10、30、50 的时候，当主题数目为 15 时，算法 MAE 最小，表明不管在何种近邻数目下，设置主题数目为 15，更加有利于保证算法的推荐效果。因此，在算法中主题数目的最佳值为 15。

#### 3.3.3 权重系数 $\lambda$ 的确定

为了确定最佳的权重系数 $\lambda$，分别设置近邻数目为 10、30 和 50，通过调节权重系数 $\lambda$，计算 IUCF 算法在不同 $\lambda$ 值下的 MAE，如图 4 所示。

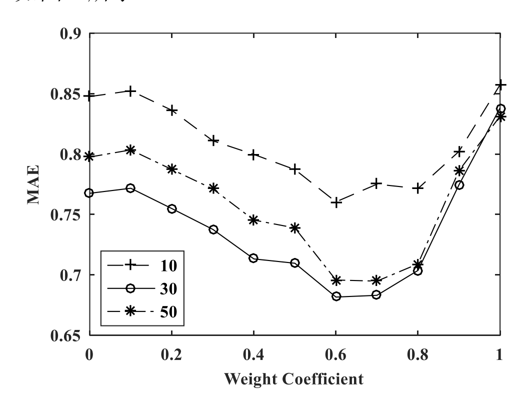



图 4　不同权重系数下的 MAE 值变化

从图 4 可以看出，对于近邻数目为 10、30、50，当 $\lambda$=0.6 时，IUCF 算法的 MAE 最小，算法的推荐质量最佳。因此，实验中权重系数 $\lambda$ 值取为 0.6。

### 3.4　算法对比实验

为了充分比较 UCF 算法与其它算法在推荐准确度上的差别，分别选择 4 种算法作为对比算法进行实验，包括传统的于用户的协同过滤算法（UCF）和基于项目的协同过滤推荐算法（ICF），文献[4]所提出的基于特征转移和概率矩阵分解的推荐算法（FTMF），文献[11]所提出的采用信任网络增强的协同过滤算法（ECFATN），实验结果如图 5 所示。

图 5 中，相较于传统的 UCF 和 ICF，IUCF 在 MAE 上明显下降，表明本文算法确实提高了传统协同过滤算法的准确度。对比 ECFATN，在不同近邻数目下本文算法 MAE 更优，表明本文算法更优，但是对比 FTMF，当近邻数目增加到 30 以后，IUCF 的 MAE 要高于 FTMF，这可能是由于项目偏差分中过多的近邻数目反而降低了算法的准确度，但是在近邻数目少于 30

时，IUCF 的 MAE 要低于 FTMF，表明本文算法在近邻数目较低时，提高了算法的准确度。

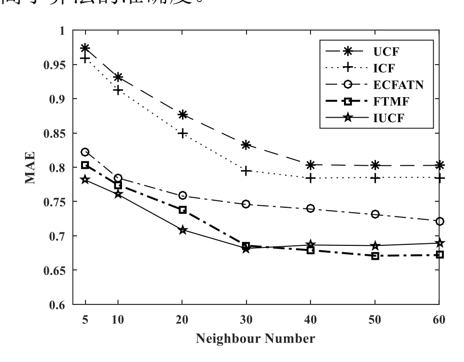



图 5 不同算法下的 MAE 值变化

为了验证 IUCF 算法与其他算法在时间效率上的差别，分别选择 UCF 和 ICF 进行对比实验，分别对 20、40、60 名用户产生推荐，不同算法的运行时间对比如表 1 所示。

表 1 算法运行时间对比表 /s

| 用户数 | UCF | ICF | IUCF |
|---|---|---|---|
| 20 | 3.79 | 4.15 | 4.03 |
| 40 | 6.12 | 7.82 | 6.92 |
| 60 | 10.87 | 12.54 | 11.32 |

从表 1 中可以看出，IUCF 算法在时间效率上要明显优于传统的基于项目的协同过滤推荐算法，但是与基于用户的协同过滤推荐算法相比，IUCF 算法的运行时间较长，这是因为基于项目的协同过滤相似度的计算要在线完成，而基于用户的协同过滤的相似度计算可以离线完成。

综上，IUCF 算法在推荐准确度和时间效率上得到了一定的改善。

## 4 结束语

准确衡量项目偏差与用户偏好对于推荐质量有较大影响，通过限定计算范围，排除不相关的项目和用户，有效地缓解协同过滤推荐中项目和用户间关联因素的相互影响导致的准确度不高的问题，取得了较好的效果。

下一步可以采用自然语言处理技术，从用户的评论中提取情感词，进而挖掘项目偏差与用户偏好，进一步提高推荐系统的准确度，同时使推荐结果更具有说服力。

## 参考文献：


[1] Chen Wei, Hsu W, Lee M L. A unified framework for recommendations based on quaternary semantic analysis [C]// Proc of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. New York: ACM Press, 2011: 1023-1032.

[2] Shi Min, Liu Jianxun, Zhou Dong, *et al.* A probabilistic topic model for mashup tag recommendation [C]// Proc of IEEE International Conference on Web Services. 2016: 444-451.

[3] Qiao Zhi, Zhang Peng, He Jing, *et al.* Combining geographical information of users and content of items for accurate rating prediction [C]// Proc of the 23rd International Conference on World Wide Web. New York: ACM Press, 2014: 361-362.

[4] Zhao Zhilin, Wang Changdong, Wan Yuanyu, *et al.* FTMF: recommendation in social network with feature transfer and probabilistic matrix factorization [C]// Proc of International Joint Conference on Neural Networks. 2016: 847-854.

[5] Zhou Xiuze, Wu Shunxiang. Rating LDA model for collaborative filtering [J]. Knowledge-Based Systems, 2016, 110: 135-143.

[6] 原福永，温志慧，梁顺攀，等．融合项目分类的加权 Slope One 算法 [J]. 小型微型计算机系统，2017，38 (09): 2090-2095. (Yuan Fuyong, Wen Zhihui, Liang Shunpan, *et al.* Integrating Item Category Into Weighted Slope One Algorithm [J]. Journal of Chinese Mini-Micro Computer Systems, 2017, 38 (09): 2090-2095. )

[7] 刘慧婷，陈艳，肖慧慧．基于用户偏好的矩阵分解推荐算法 [J]. 计算机应用，2015，35 (S2): 118-121. (Liu Huiting, Chen Yan, Xiao Huihui. Matrix factorization recommendation algorithm based on users’ preference [J]. Journal of Computer Applications, 2015, 35 (S2): 118-121. )

[8] Herlocker J L, Konstan J A, Borchers A, *et al.* An algorithmic framework for performing collaborative filtering [C]// Proc of International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1999: 230-237.

[9] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3 (3): 993-1022.

[10] Casella G, George E I. Explaining the Gibbs sampler [J]. The American Statistician, 1992, 46 (3): 167-174.

[11] 李熠晨，陈莉，石晨晨，等．采用信任网络增强的协同过滤算法 [J]. 计算机应用研究，2018，35 (01): 116-120. (Li Yichen, Chen Li, Shi Chenchen, *et al.* Enhanced collaborative filtering adopting trust network [J]. Application Research of Computers, 2018, 35 (01): 116-120. )

[12] 何明，孙望，肖润，等．一种融合聚类与用户兴趣偏好的协同过滤推荐算法 [J]. 计算机科学，2017，44 (S2): 391-396. (He Ming, Sun Wang, Xiao Run, *et al.* Collaborative filtering recommendation algorithm combining clustering and user preferences [J]. Computer Science, 2017, 44 (S2): 391-396. )

[13] 罗淳．关于人口年龄组的重新划分及其蕴意 [J]. 人口研究，2017，41 (05): 16-25. (Luo Chun. Re-partitioning population age group and its implications [J]. Population Research, 2017, 41 (05): 16-25. )

[14] Wood J C, Wood M C. Joseph M. Juran: Critical evaluations in business and management [M]. [S. l. ] : Psychology Press, 2005.

[15] Bholowalia P, Kumar A. EBK-means: A clustering technique based on elbow method and k-means in WSN [J]. International Journal of Computer Applications, 2014, 105 (9) .

[16] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National academy of Sciences, 2004, 101 (suppl 1): 5228-5235.